International Environmental Modelling and Software Society (iEMSs)
7th International Congress on Environmental Modelling and Software
San Diego, California, USA, D.P. Ames, N. Quinn (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

The data processing inequality and environmental model prediction

Steven V. Weijs [a]

[a] *Ecole Polytechnique Federale de Lausanne, Laboratory of Cryospheric Sciences Station 2, 1015, Lausanne, Switzerland (steven.weijs@epfl.ch)*

Abstract: Prediction in environmental systems, such as hydrological streamflow prediction, is a challenging task. Although on a small scale, many of the physical processes are well described, accurate predictions of macroscopical (e.g. catchment scale) behavior with a bottom-up mechanistic approach often remains elusive. On the other hand, conceptual or purely statistical models fitted to data often perform surprisingly well for prediction. The data processing inequality, from the field of information theory, says that processing data with statistical procedures can only decrease, and not increase the information content of the data. This seems to contradict the intuition that our knowledge of physical processes should help in making informed predictions with simulation models fed by environmental data. In this paper, we propose a perspective from information theory and algorithmic information theory, to resolve this apparent contradiction and to shed light on where the information in environmental predictions originates from. Algorithmic information theory relates information content to description length and therefore enables an intuitive view of inference as a form of data compression, in which information in data is compactly represented by the patterns that can be discovered in it.

Keywords: hydrology, physically based modeling, algorithmic information theory, model complexity.

1 INTRODUCTION

In hydrology, like other environmental sciences and science in general, the goal is to find descriptions of reality that enable testable predictions. Since most systems in our natural environment are complex and incompletely observed, these predictions will most likely contain significant uncertainties. These uncertainties should ideally be represented in the prediction in terms of probability. Probabilistic predictions, as argued by Weijs et al. [2010a], enable unambiguous evaluation in terms of information and uncertainty, which are measurable quantities in the framework of information theory, developed and described by Shannon [1948]; Cover and Thomas [2006].

Information-theoretical evaluation of probabilistic predictions against observations, using the framework of Weijs et al. [2010b], which was extended to the case of uncertain observations by Weijs and Van de Giesen [2011], enables us to measure the information content of predictions. This information content originates from the mutual information, see Cover and Thomas [2006], between predictor variables and the variable to be predicted, and is reduced by wrong information due to miscalibrated probability estimates.

Prediction in the environmental sciences can thus be seen as an attempt to channel information flows to the prediction. This is achieved through 1) tapping the most important sources of information by observation of the environment and 2) using prior knowledge of

physical processes to build models that incorporate this information and transform it into informative predictions and 3) using inference, model calibration and data assimilation to extract information from observations and channel it to model predictions.

The question can then be asked how much of the information present in the predictions originates from the knowledge of physical processes and how much of the information originates from the data supplied to a specific model.

Data processing inequality, one of the (non-)conservation "laws" of information theory, states that data processing can never increase, but only decrease, the amount of information in the data [Cover and Thomas, 2006]. This seems to contradict the intuition that our knowledge of physical processes should help in making informed predictions with simulation models fed by environmental data.

In this paper, we use a perspective from information theory and algorithmic information theory, to resolve this apparent contradiction and to shed light on where the information in environmental predictions originates from.

1.1 Characteristics of prediction in environmental systems

Environmental systems are often characterized by significant complexity and limited observability. The complexity originates for example from a large number of (hidden) states, non-linear behaviour, and feedbacks at a wide range of interacting scales. Knowledge of physical processes is typically about the small (lab-)scale, while the scale of interest are often larger, such as for example the catchment scale in hydrology. Simulation models follow a physically based, bottom up approach to prediction, where physical processes are modeled on the small scale in a distributed manner, while behaviour on scales smaller than the model scale is parametrized. These models are considered to have the largest degree of realism. However, the emergent macroscopic behaviour of complex environmental systems is often hard to predict bottom up, due to insufficient characterization of the dynamics and lack of data. Simple conceptual or empirical models fitted directly to the macroscopical behaviour on the scale of prediction often perform remarkably well. Gupta et al. [2012] compare the attitudes to modeling in different fields deal with the terrestrial hydrosphere.

1.2 Problem

Both the data driven (top-down) and physically-based (bottom-up) modeling approaches are useful for making predictions and understanding the systems on different scales, but combining the strengths of both approaches remains difficult. Although in theory knowledge of physical processes should result in better predictions, there are many problems plaguing distributed physically based hydrological modeling [Beven and Binley, 1992; Grayson et al., 1992; Beven, 2001]. At the same time, there is little hope that the important questions that environmental science faces can be answered with simple conceptual models or data driven models alone, especially in the context of predictions under change [Ehret et al., 2014].

1.3 Solution: information as central concept in environmental science

The resulting core problem in hydrological prediction is therefore to channel the flows of information, resulting from knowledge of microscopic physical processes, macroscopic

organizing principles, and observations on widely varying scales, so that it percolates into the forecasts and decisions, while maximally avoiding the pollution by misinformation, resulting from overconfidence in theories or measurement uncertainties that are not accounted for.

Since also the knowledge on small scales and organizing principles eventually originates from observation, which is extraction of information from the environment through sensors, we can depict the information flows in environmental science in the simplified scheme in Figure 1. The lower two blocks show how information obtains value trough decisions, which enable information to flow back into the environment to make it more useful to some specified needs of groups or individuals. This is where environmental science becomes environmental engineering.

The success of our scientific efforts can be measured by the information content of our predictions and the success of engineering can be expressed in the form of the utility added to our environment. When the information flows on the red arrows can be quantified, this offers the possibly to approach the challenge of optimized environmental predictions in a more structured way. Information theory offers that potential.

Optimizing information flows also needs detailed information-theoretical analysis of the intermediate arrows, such as quantification of the information content of data and model complexity. To find universal approaches to reach optimal predictions in different situations, the intricate dynamics of information should be fully understood.

This paper attempts shed light on one part of these dynamics: the role of the data processing inequality. These initial thoughts will hopefully serve as a starting point for discussion that contributes to better understanding of information flows.

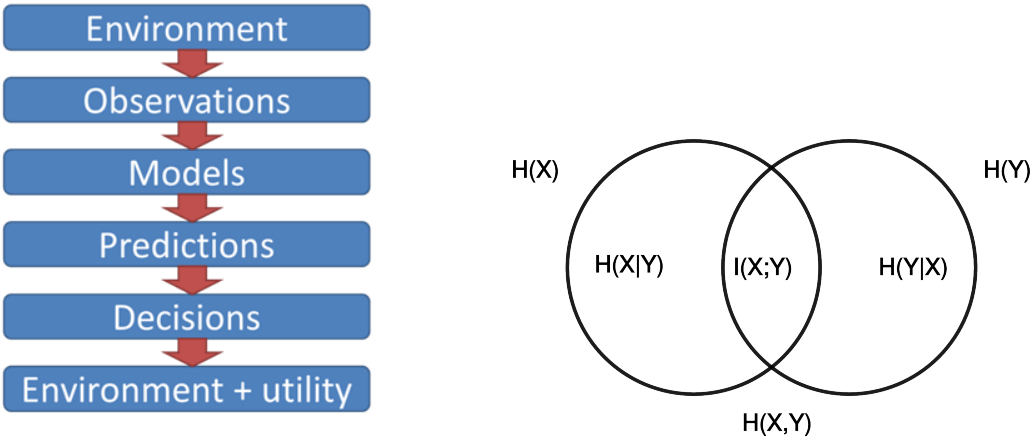

Figure 1. Left: Simplified conceptualization of information as a flow. Right: Additive relations between information-theoretical quantities.

2 Background

2.1 Information theory

Shannon [1948] laid the foundation for the field of information theory by defining a mea-

sure for uncertainty, or missing information, named entropy, defined as a function of a probability distribution, which uniquely satisfies some basic requirements and additive properties. Information theory also defines conditional entropy, $H(X|Y)$, (the uncertainty left in X when knowing Y) and mutual information, $I(X;Y)$, (the reduction of uncertainty in Y, given knowledge of X); see Figure 1. These measures are defined in eq. 1 and 2, where $p(x_i)$ stands for the probability that X takes the i^{th} possible discrete outcome x_i.

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i), \;\; H(X|Y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(y_j)}{p(x_i, y_j)}, \tag{1}$$

$$I(X;Y) = \sum_{i,j} p(x_i, y_j) \log \left(\frac{p(x_i, y_j)}{p(x_i)\, p(y_j)} \right) \tag{2}$$

2.2 The data processing inequality

Informally, the data processing inequality states that signal processing of a random variable can never increase information content about a related random variable. Formally, if $X \leftrightarrow Y \leftrightarrow Z$ form a Markov chain, meaning that Z is conditionally independent of X, given Y. Then the mutual information between X and Z can never be larger than the mutual information between X and Y or between Y and Z;

$$I(X;Y) \geq I(X;Z) \text{ and } I(Y;Z) \geq I(X;Z).$$

In the context of model prediction, this result seems to be applicable to the Markov chain formed by predictand $\leftrightarrow$ predictor $\leftrightarrow$ prediction; see also Gong et al. [2013]. The prediction is conditionally independent of the predictand, given the predictor, since the predictor fully determines the prediction and predictand does not contribute additional information.

2.3 Algorithmic information theory (AIT)

Because information theory defines information content in terms of probabilities, the question is how to assign probabilities to various observed or potentially observable outcomes. This can be problematic, especially if little data is available. Also intuitively it is unsatisfying that the information content of observed data depends on what one assumes could have been observed instead. Algorithmic information theory bypasses these issues by defining information content of data as the shortest computer program on a reference computer that can generate those data, or more loosely as the shortest description of the data in a reference language. Therefore it can assign an information content (and probability!) to a single object. It was found that building on this definition, a theory could be built that is analogous, and in the limit equivalent to Shannon's definition of information. Since there is no space to go into more details here, the interested reader referred to Li and Vitanyi [2008] and references therein.

3 PERSPECTIVE ON ENVIRONMENTAL MODEL PREDICTION

3.1 Example: data processing inequality for a hydrological model

Consider the prediction of streamflow from meteorological variables. For simplicity, we assume that the only data available are rainfall and observed streamflow. On the one hand, the data processing inequality (DPI) could be interpreted as saying that whatever structure is put into the model, the predictions can never get better than dictated by the mutual information between the input data and predicted variable. On the other hand, it is widely believed that more knowledge of the underlying equations describing the physics will lead to improved predictions.

A first way in which these two ideas can be reconciled is by assuming that practically, all models lose information from the input data, and that there is always something to gain by finding a model that loses less information. The physical understanding could then help us selecting the right model. However, this also implies that we should be able to find this same model without having physical knowledge by simply trying different black box models until we find the one that retains most information from the predictors about the predictand.

The second way in which the value of process knowledge and the DPI can be reconciled is seeing models as possible containers of information, rather than merely a transformation of the input variables to the output variables. Physical laws implemented in the model code can be seen as a highly condensed form of information from previous observations that were used to infer that law. Informative past data is thus channeled into the model through the model equations and therefore the DPI is respected, even if the model seems to add predictive power to the input data.

According to Bayes law, the relative importance of the prior information fades away when more data become available. We therefore can expect that model complexity and information content of data should play a role in the dynamics of information flow through a model.

3.2 AIT view on information content in data and model complexity

Recently, attempts have been made by Gong et al. [2013] to quantify information content in data , thereby quantifying the maximum possible model performance. It is important to notice that information content of data is always subject to a model, which depends on prior knowledge; see for example Weijs et al. [2013a, b]. This prior knowledge is relatively more important when little data available, since for large amounts of data, the probabilities of the different outcomes converge to the observed frequencies.

In the limit of very long time series of inputs and outputs, the added knowledge of physics put into the model loses importance, and the predictive information that the input data contains about the output data is dominated by the mutual information. It is in these cases that the data processing inequality applies. If if the model has high complexity (has a long shortest description) compared to the input data, then the complexity of the predictions could theoretically be higher than that of the predictors, so in the AIT view, a model could add information to the input data, and if that information is correct, it could actually increase mutual information with the observed output. However, deterministic processing cannot increase the information content by more than a constant; see Grunwald and Vitányi [2004]. In other words, models can add information, but just a constant amount that becomes less important for longer time series.

4 CONCLUSIONS

From the discussion in the previous section it can be concluded that there are various subtleties to be considered when applying the data processing inequality to environmental model prediction. Tentatively, it can be concluded that the inequality holds also for models reflecting process understanding, but probably needs to be formulated in terms of algorithmic information theory. Especially when there is relatively little data, high model complexity that is warranted by actual knowledge of physics can actually contribute to informativeness of predictions, beyond what is possible from the data alone.

It is important to realize that all knowledge about our environment ultimately stems from observation. The DPI tells us that, eventually, new observations are needed to improve environmental science. Models giving right answers for wrong reasons [Kirchner, 2006] is nothing more than models giving some wrong answers next to the right answers. Therefore, the way forward to obtain the right answer for the right reasons is to ask our models more questions, and collect more observations to check the answers given by the model.

4.1 Making testable predictions: an art or a procedure?

Understanding of our environment is achieved by making - and results in - testable predictions. Whether or not physically based distributed modeling of the environment is to be preferred over simple conceptual or empirical models that are inferred from the data depends entirely on the specific case. Specifically, the case is determined by the information in the available data, the relevant knowledge of physics and emergent behaviour on all scales, and the quantity to be predicted.

One could stop there and conclude that environmental modeling is an art, where the modeler / artist uses creativity to decide how to approach the modeling task [Savenije, 2009]. However, when the full dynamics of information content and flows are properly understood, there is potential to replace part of this creative freedom with optimal procedures that lead to improved predictions. Currently however, human creativity and pattern recognition capabilities are unsurpassed by completely mechanical procedures ("effective methods") for inference in the practice of environmental sciences. Juxtaposing both human intelligence and artificial intelligence in the environmental sciences will likely improve both sides and especially their combination. Therefore we should continue to learn from attempting to formalize and quantify the process of learning from data in the environmental sciences.

ACKNOWLEDGMENTS

The author wishes to thank Hoshin Gupta, Grey Nearing and Wei Gong for interesting discussions that helped form the ideas presented in this paper and the AXA Research Fund for financial support.

REFERENCES

Beven, K. J. (2001). How far can we go in distributed hydrological modelling? *Hydrology and Earth System Sciences*, 5(1):1–12.

Beven, K. J. and Binley, A. M. (1992). The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes*, 6:279–298.

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory*. Wiley-Interscience, New York.

Ehret, U., Gupta, H. V., Sivapalan, M., Weijs, S. V., Schymanski, S. J., Blöschl, G., Gelfan, A. N., Harman, C., Kleidon, A., Bogaard, T. A., Wang, D., Wagener, T., Scherer, U., Zehe, E., Bierkens, M. F. P., Di Baldassarre, G., Parajka, J., van Beek, L. P. H., van Griensven, A., Westhoff, M. C., and Winsemius, H. C. (2014). Advancing catchment hydrology to deal with predictions under change. *Hydrology and Earth System Sciences*, 18(2):649–671.

Gong, W., Gupta, H. V., Yang, D., Sricharan, K., and Hero, A. O. (2013). Estimating epistemic & aleatory uncertainties during hydrologic modeling: An information theoretic approach. *Water Resources Research*, page in press.

Grayson, R. B., Moore, I. D., and Mc MAHON, T. A. (1992). Physically based hydrologic modeling 2. is the concept realistic. *Water Resources Research*, 26(10):2659–2666.

Grunwald, P. and Vitányi, P. (2004). Shannon information and kolmogorov complexity. *arXiv preprint cs/0410002*.

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., and Ye, M. (2012). Towards a comprehensive assessment of model structural adequacy. *Water Resources Research*, 48(8).

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, 42(3).

Li, M. and Vitanyi, P. M. B. (2008). *An introduction to Kolmogorov complexity and its applications*. Springer-Verlag New York Inc.

Savenije, G. (2009). HESS Opinions: 'The art of hydrology'. *Hydrology and Earth System Sciences*, 13(2):157–161.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical J.*, 27(3):379–423.

Weijs, S. V., Schoups, G., and van de Giesen, N. (2010a). Why hydrological predictions should be evaluated using information theory. *Hydrology and Earth System Sciences*, 14(12):2545–2558.

Weijs, S. V. and Van de Giesen, N. (2011). Accounting for observational uncertainty in forecast verification: an information–theoretical view on forecasts, observations and truth. *Monthly Weather Review*, 139(7):2156–2162.

Weijs, S. V., van de Giesen, N., and Parlange, M. B. (2013a). Data compression to define information content of hydrological time series. *Hydrology and Earth System Sciences*, 17(8):3171–3187.

Weijs, S. V., van de Giesen, N., and Parlange, M. B. (2013b). Hydrozip: How hydrological knowledge can be used to improve compression of hydrological data. *Entropy*, 15(4):1289–1310.

Weijs, S. V., Van Nooijen, R., and Van de Giesen, N. (2010b). Kullback–Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition. *Monthly Weather Review*, 138(9):3387–3399.